MASTER'S THESIS

The importance of subject matter experts for the success of managing data science projects

Simons, R (Robert)

Award date:
2020

The importance of subject matter experts for the success of managing data science projects

Opleiding:	Open Universiteit, faculteit Management, Science & Technology Masteropleiding Business Process Management & IT
Programme:	Open University of the Netherlands, faculty of Management, Science & Technology Master Business Process Management & IT
Cursus:	IM0602 Voorbereiden Afstuderen BPMIT IM9806 Afstudeeropdracht Business Process Management and IT
Student:	Robert Simons
Identiteitsnummer:	
Datum:	09-07-2020
Afstudeerbegeleider	Prof.dr.ir. Remko Helms
Meelezer	Dr.ir. Harry Martin
Versie nummer:1.0	
Status:	FINAL

Abstract

The purpose of this thesis is to get a better understanding of the knowledge gap on the challenges and success factors (C&SF) that influence the success of managing data science projects (DSP) and their relationship to each other via exploratory qualitative research. This resulted in 42 unique C&SF's, which can be summarized in to six major categories: *Value*, *People*, *Technology*, *Data*, *Process* and *Organization*. These six major categories are used in the preparation for the interview protocol and in the Thematic Analysis procedure for the data analysis. An embedded multiple exploratory case study was done with three companies. These three cases studies were explored on several DSP's related topics and shared many similarities and differences. Via cross case comparison all these relevant interrelated findings have led to the result of seeing a strong relationship of these findings in combination with the importance of the C&SF of having a subject matter expert (SME) actively participating in DSP's.

Recommendations will be made on how these findings can be best used in practice and suggestions are made for future research on this topic.

Key words

Data Science Project (DSP), Subject Matter Experts (SME), Case Studies, challenges and success factors (C&SF)

Summary

The purpose of this thesis is to get a better understanding of the knowledge gap on the challenges and success factors (C&SF) that influence the success of managing data science projects (DSP) and their relation to each other. The first chapter will provide some background information about the management of data science projects and explain the problem that goes along with that management. A systematic literature review was done to gather and assess high-quality literature relevant to this research. This was analysed and provided a list of 42 unique C&SF's that influence the success of managing DSP's, which can be summarized in to six major categories: *Value*, *People*, *Technology*, *Data*, *Process* and *Organization*. These six major categories will be explained one by one and are used in the preparation for the interview protocol and in the analysis of the case studies.

For this research there was a need to get an understanding of how companies are managing DSP's in the practice. To get that information an embedded multiple exploratory case study was done with three companies. By using the information gained from the literature study as "a priori" codes it was possible to do the data analysis via a Thematic Analysis procedure. This allowed the discovery of a relationship between the subject matter expert (SME) C&SF and a multitude of other C&SF's.

Three companies were chosen for this study. They stated that they have 10+ years of experience in DSP's, which makes them experienced. These three cases will be explored on several DSP related topics like *"Strategy, Organizational setup, Mindset, Approach, etc."*, which via a cross case comparison will highlight the biggest differences and similarities between these case studies. The differences are their resource handling and focus of their specialized data science team in managing DSP's. Twelve profound C&SF's will be discussed, which are all recognized and to a certain extent similar among these three case studies.

All of these relevant interrelated findings suggested that there is a strong relationship between the success of DSP and having a subject matter expert (SME) actively participating in DSP. There is a relationship found in six of the twelve profound C&SF's for these case studies. For these six C&SF's it was found that the SME could have a dependency influence or accelerator capability towards the related C&SF's. The type of effect will be explained for having a SME actively participating in DSP's versus the other impacted factors. The other contributing "solutions" of the cases studies will also be highlighted in their relation to one and other.

Recommendations will be made on how these findings can be best used in practice and suggestions are made for future research on this topic.

Introduction

1.1. Background information

Data science might sound new but it is at least more than two decades old (Saltz, Shamshurin, & Crowston, 2017). Data science is an enabler to improve the techniques to understand (big) data. It is already so embedded in the way the world works that most people are not even aware that this is for a great part done via data science. Obvious examples are weather forecasts or our navigation tools like TomTom or Google Maps (Moya-Gómez, Salas-Olmedo, García-Palomares, & Gutiérrez, 2018; Philip Chen & Zhang, 2014), but also the way we do shopping offline and online, benefits from data science. Webshops can come up with proposals of items for you to buy based upon a profile they were able to design for you personally. The biggest companies in the world at this moment Google, Amazon, Facebook and Alibaba Group so heavily depend on data science, they would not have existed or would no longer exist if they would not have applied data science (MGI, 2016).
There is no best practice on applying data science and there is even less focus on the management of a data science project(DSP) (Saltz et al., 2017). Not all DSP's are successful, according to some that is even an understatement and that between 60-85% of all DSP's fail (MGI, 2016; Walker, 2017). This is due to companies not being mature enough in managing DSP's.
Because data science is IT-related, people might think we can simply apply a software development cycle and manage a DSP like a software project. This approach can really backfire because data science is more like research than it is like engineering (Provost, 2013).
There is a research from Saltz et al. (2017) on "comparing data science project management methodologies via a controlled experiment", there are also a lot of other process-, people- and technology-related factors that influence DSP (Koronios, 2015). However, no consensus exists in literature on how to manage DSP's.

1.2. Problem statement

Understanding how to manage DSP's in such a way that they improve business results is a general problem. Organizations might have managerial problems with providing enough time for DSP's or it can even be a cultural barrier, which has to be dealt with before an organization is ready for data science (LaValle, Lesser, Shockley, Hopkins, & Kruschwitz, 2011). Data science projects are influenced on their success by several factors. Data science projects are not always successful and require a lot of experimenting, it does not come as a surprise to see that DSP managers are more successful if they are able to adapt and be more flexible in their planning, compared to the more classic IT project managers. In order to do that, they need to have a good understanding of data science experiments and its related processes and see the added value in that (Viaene, 2011).

Summarizing the problem statement:
Companies are failing to manage data science projects successfully.

1.3. Research question

There are many C&SF's influencing the success of DSP's. These C&SF's vary from being within the perception of business to the structure and properties of the data itself. They all have a certain level of influence on the outcome of a DSP. Providing an overview of these C&SF's that influence the success of DSP's can help companies improve their success rate of those projects.

Research question:

How can a company ensure the successful execution of a data science project (DSP)?

a. *What are the minimal requirements for doing a DSP within a company (process, people (roles and skills), data, etc.)?*

b. *How can a company best set-up their organization to do DSP's?*

To ensure the successful execution of DSP's a company has to make sure they correctly set-up their internal organisation who is dealing with DSP's and also make sure that the minimal requirements like process, people, data, etc. are up to par. By answering these sub questions companies should be able to reflect to their own setup or plan for doing DSP's

1.4. Relevance

There is currently a gap in knowledge on the C&SF's of manage DSP's. For example, which project methodology works best for managing a DSP or what would the influence be of the setup of the company's IT-organization, which should support DSP's (Obasa, Salim, & Al-Taie, 2017; Saltz et al., 2017; Vidgen, Shaw, & Grant, 2017).
From a scientific perspective, this would provide insights on what is important for managing DSP's and what C&SF's and/or requirements are important when a company is doing DSP's. For the social importance, this research could provide a basic guide for organizations to do successful DSP's (Matsudaira, 2015; Stimmel, 2015; Vidgen et al., 2017). By using this research companies could recognizing potential pitfalls they might encounter during DSP's or already have present within their organization, this will allow them to take action sooner, which could ensure the success of DSP's or just save costs by cancelling them or putting them on hold.

1.5. Research approach

The methodological approach was an exploratory qualitative research because we would like to understand the C&SF's and their relations that have influence on successfully managing DSP's.
This scientific literature delivers a first full overview of all relevant C&SF's that come into play with managing DSP's.
The overview was subdivided into the six most important categories, which have been found during the literature research. Based on those categories there were a total of nine semi-structured interviews done at three different "scientific" top performing companies that are doing DSP's, in which the completeness of the overview and the importance of all factors was validated or enriched.

2. Literature review

2.1. Literature Research Approach

Okoli & Schabram (2010) have created an eight-step guide in conducting a systematic literature review of which three are applied in this research; *Searching for the literature, Practical screen Quality Appraisal, Data extraction and Synthesis of studies* (Okoli & Schabram, 2010) (see table 1.).

Searching for the literature

It is important to be explicit in what is searched for in the literature and in what is considered useful for this research. It is important that people can understand that a comprehensive search of literature was done. Data science is a subject, which is mostly researched in the interest areas of "Computer Science", "Business" and "Library & Information Science", to keep the scope strict focus will be on these, which are peer-reviewed and recent studies, preferably published in highly-prestigious journals (Knopf, 2006). A start set of literature was available, which shapes the basis for this research.

Practical screen

By specifically searching for papers that describe real-world information about managing DSP's and its related factors were considered for this research. This was achieved via the means of filtering within search queries.

Quality Appraisal

Literature found and considered matching the research was also screened for exclusion, based on a quality scoring, which is dependent on the research methodology applied to those articles (Okoli & Schabram, 2010).

Data extraction

The literature resulting from the quality appraisal is considered valuable and from that the applicable information will be exacted from each study.

Synthesis of studies

In the analysis the insights per study will be combined on similar basis like that of a study that is considered the most important, via a combination of qualitative and quantitative techniques.

2.2. Literature Research Execution

Having high-quality literature (Knopf, 2006) related to the research question was the first step of this research, this was achievable via the use of the university's online scientific library. First step in the *Searching for the literature* was getting the terminology right ("Data Science" can also mentioned differently in literature like "Data Mining" or "Data Analytics".), which was accomplished via a custom search query (table 1). The research query was enhanced with "filters" in the sections of *Review*, *Interest areas*, *Material* and *Term* selections. This enhanced query resulted in a list of 561 documents, which were input for the *screening* (Okoli & Schabram, 2010). Screening was done by searching for topics that come close to the research topic and filter those that are not related to the topic (Okoli & Schabram, 2010). The *screening* resulted in 46 papers (on top off the starter set of seven), which were then used for the *quality appraisal* (Okoli & Schabram, 2010). The end result is fourteen papers, which are considered to be of high enough quality and contain useful information for this research.

Table 1 | Literature search evolution

Phase	Query	Filters		results
Searching for the literature	(((TitleCombined:("Data Science")) OR ("Data Analytics") OR ("Data Mining")) AND (("projects") OR ("Manag*")))))	Main filters:	No articles older than 2015, Peer-Review only, Scholarly & Peer-Review	561
		Interest Area:	"Computer Science", "Business" and "Library & Information Science".	
		Material selection:	Magazine, Publication, eJournal, Manuscripts	
		Terms Selection:	Data mining, Big data, Data mining and knowledge discovery, Information management, Humans, Data management, Information, Management, Decision making, Research article	
Practical screen	Go through all the papers and filter out papers that did not fit my research.	Manual activity: • Only select papers that describe parts about managing data science projects and the "challenges" it faced or important steps it took for success or "lessons learned" Papers that for example describe how to best setup a data clustering algorithm is not in my scope.		53
Quality appraisal	Go through all the papers and validate them.	Manual activity: • Validate that the authors are credible. • Validate that the sources used in the papers are credible		14

From the literature research the paper of Vidgen (2017) was the most important and came closest to this research, his main list of "Management Challenges" was taken as a fundament, this was enriched from other literature research towards 42 unique C&SF's that influence the success of managing DSP's. All of these combined findings are spread across 154 sub C&SF's (See appendix 1.1 and 1.2) and can be traced back to six major categories as described by Vidgen (2017): *Value*, *People*, *Technology*, *Data*, *Process* and *Organization*.

2.3. Results and conclusions from the literature research: six major categories explained

Value

Insights generated by Data Science Projects (DSP's) should provide value for business. That is if the business is able to use their analytics to justify, guide and prescribe their actions to do fact-based decision-making (LaValle et al., 2011; Vidgen et al., 2017). It all starts with making sure you have a well-established business case, which supports a clearly defined business goal, where the perceived costs are lower than projected benefits (Dutta & Bose, 2015; Koronios, 2015; LaValle et al., 2011; Schüritz, Brand, Satzger, & Bischhoffshausen, 2017; Schwarz, Schwarz, & Black, 2014; Viaene, 2011; Vidgen et al., 2017). It is important that the business is aware of the value of data science and consider it as a strategic instrument to be used as an enabler for their future strategy and at the same time optimize their day-to-day operations (Dutta & Bose, 2015; Koronios, 2015; LaValle et al., 2011; Philip Chen & Zhang, 2014; Schüritz et al., 2017; Schwarz et al., 2014). A good way of showing the value is by measuring the customer value impact and making sure that the expectations of the business are managed correctly and the analytical results are translated into a format that is easily understandable for the business (Koronios, 2015; Vidgen et al., 2017).

People

Companies have their organizational unit who responsible for DSP configured as an independent, centralized analytical unit, which is in line with the business analytical units, but not part of the IT-department (Koronios, 2015; LaValle et al., 2011; Schüritz et al., 2017). For the unit doing the DSP's, it is important to make sure that the analytical and technical skills of not only the data scientists but also the business are of a high enough level. This is imperative to avoid the knowledge gap of how to benefit from data science to improve their business (LaValle et al., 2011; Obasa et al., 2017; Schwarz et al., 2014; Viaene, 2011; Vidgen et al., 2017) and set-up these units with multidisciplinary, cross-functional teams, which have a great diversity in their setup (Dutta & Bose, 2015; Koronios, 2015; Obasa et al., 2017).

Technology

Existing IT platforms are considered to be a limiting factor, unless these have a strong data infrastructure with a standardized tools selection that is compatible with the companies legacy systems, so that it enables resource-sharing and reduces maintenance and licensing costs (Dutta & Bose, 2015; LaValle et al., 2011; Schwarz et al., 2014; Stimmel, 2015; Viaene, 2011; Vidgen et al., 2017). New analytical platforms, normally have an end-to-end solution like an Analytics-as-a-Service setup with the correct analytical tools that are able to manage high data volumes and are supported by security-, control versioning-, managing load and performance-capabilities (Koronios, 2015; Philip Chen & Zhang, 2014; Schüritz et al., 2017; Schwarz et al., 2014; Stimmel, 2015; Vidgen et al., 2017).

Data

Data is the most mentioned category where you can find influencing factors within the literature (see table 2 in the appendix). One of the reasons for that is that you should keep in mind the full data management life-cycle, but only after you have determined if the data is available, accessible, of high enough quality and useful (Koronios, 2015; LaValle et al., 2011; Philip Chen & Zhang, 2014; Schüritz et al., 2017; Viaene, 2011; Vidgen et al., 2017). Then you need to be able to integrate and migrate the data keeping in mind the technical, security and privacy requirements (Koronios, 2015; Schüritz et al., 2017; Schwarz et al., 2014; Viaene, 2011; Vidgen et al., 2017). Once your DSP's have created insights a setup must be made about that newly generated data, like who is allowed to access this via what methods and how can this data be visualized for your business so they can gain the benefits (Koronios, 2015; LaValle et al., 2011; Obasa et al., 2017; Philip Chen & Zhang, 2014; Schüritz et al., 2017; Schwarz et al., 2014; Stimmel, 2015; Viaene, 2011; Vidgen et al., 2017).

Process

At a high-level, a good data governance setup should be available that allows prioritization on master data sources and the rework to capture enterprise efficiencies (LaValle et al., 2011; Schüritz et al., 2017). It is advised to have a well-defined project management setup following a clear methodology, which allows intelligent experimentation for doing DSP's within an accelerated process(Koronios, 2015; Saltz et al., 2017; Viaene, 2011; Vidgen et al., 2017). Sharing all your work via knowledge management is highly recommended because that will broaden the adoption of analytics within the company (Dutta & Bose, 2015; Koronios, 2015; Matsudaira, 2015; Schüritz et al., 2017).

Organization

Ensure that data science is part of the organization's strategy as this would allow you to build a corporate data culture within the organization so you can change the mindset around how to deal with data (Dutta & Bose, 2015; Koronios, 2015; LaValle et al., 2011; Viaene, 2011; Vidgen et al., 2017). Have that incorporated within the organization's architecture so its becomes imbedded (Schüritz et al., 2017; Stimmel, 2015). Make sure that there is enough committed sponsorship from top management to ensure that there is time available to actually do DSP's (Koronios, 2015; LaValle et al., 2011; Schwarz et al., 2014; Vidgen et al., 2017). Assign data owners so it becomes clear who is responsible for what in regards to data and its quality, availability and accessibility (LaValle et al., 2011; Vidgen et al., 2017). Setup a strong collaboration between your IT and Business for enabling an end-to-end solution (LaValle et al., 2011; Matsudaira, 2015; Schwarz et al., 2014; Vidgen et al., 2017).

2.4. Purpose for further investigation

This scientific literature-based C&SF's list will serve as a basic guide for doing successful DSP's. Via an exploratory qualitative research this list will be validated for completeness and or will be enhanced. The six major categories and their factors are considered the most important and these are going to be validated and potentially enriched based on real world experiences of three different companies managing DSP's. Methodologies on how that is going to be done will be explained in the next chapter.

3. Methodology

3.1. Conceptual design: select the research method(s)

To best answer the research question "*How can a company ensure the successful execution of a data science project (DSP)?*" a better understanding is needed how this is done in a real-world-setting. This research must get the understanding of managing DSP's by people in the real-world, representing their views and perspectives. Get the full contextual richness of these real-world conditions, and understand that there are multiple sources of evidence (Saunders, Lewis, & Thornhill, 2016; Yilmaz, 2013; Yin, 2011). Since this research is focused on the "how" and "what" of managing DSP's, means that there is no control over behavioural events and this research will investigate a contemporary phenomenon in a real-world context, which makes this an exploratory case study. The goal is to get this understanding from three companies on how they manage DSP's, which can take place across multiple teams, getting these insights from one company will not be sufficient, because it will only provide one view while this research requires a broader understanding of how companies manage DSP's. This means this will be an embedded multiple exploratory case study, which will also allow replication on findings (Barratt, Choi, & Li, 2011; Saunders et al., 2016; Yin, 2017).
The companies relevant for this research "case" are managing / performing DSP's on a day-to-day basis. LaValle et al. (2011) points out that it is important to select top performing companies that are also active in science. This will increase the chance of those companies also managing DSP's. For this research three "scientific" top performing companies have been selected from different branches of industry to get a broader more general understanding of how DSP's are successfully managed.

3.2. Technical design: elaboration of the method

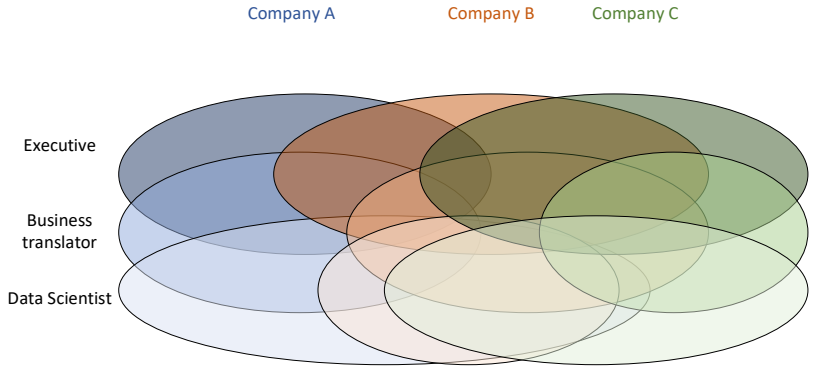
The literature research provided a list of categories and C&SF"s, which all influence managing DSP's. This list of categories will be used in the data analysis via the technique of Thematic Analysis based on a deductive approach, to better the search for themes, relations, or patterns occurring across the dataset of categories and factors (Saunders et al., 2016; Yin, 2011). These six major categories; *Value*, *People*, *Technology*, *Data*, *Process* and *Organization* will be used as 'a priori' codes for that data analysis and as a red line for creating the interviews so no important C&SF's are forgotten on what could impact DSP's (Saunders et al., 2016). Room will be left for open questions to gain insights in possible C&SF's or relations between C&SF's not yet covered within the literature. To ensure that the companies that were consulted for this information are experienced a few selection criteria are made; the companies need to work on an day-to-day basis DSP's (minimum ten projects per year) and work according a project methodology (scrum / waterfall / etc.).

Interviews are a very powerful way to collect data and understand people's views in more depth. This also allows the interviewers to communicate in their own wording and show emotions and ventilate their thoughts about the research topic, which is every useful for understanding the real factors related to this research (Alshenqeeti, 2014). The knowledge and expertise related to the six categories can be found within three different roles. Therefore, per company three persons will be interviewed, first the owner or responsible person for making sure DSP's are done, this for example could be the head of an analytics department. They can provide valuable input on how the DSP's are done from a higher managements point of view, for example the organizational setup of the team, departments involved, etc. Second person to interview is managing the DSP's, this for example could

be a project director. They will be able to provide the insights from a project management point of view, for example no resources, no commitment, etc. The last interviewee is a data scientist, which is involved in the DSP's. The data scientists can provide the information of the actual data science work and all the related factors, for example the data quality, skills, etc.

There is an overlap (the amount of overlap can variate) in roles and companies experience expected, which can be used to validate C&SF's with each other. To get a better feeling of this overlap please see an impression at figure 1.

Figure 1 | An impression of the overlap of roles and companies and their experiences



These interviews will take around one to one and a half hour and preferably take place onsite or else via a remote webcam call. All the interviews have been recorded for further analysis. As preparation for the interviews an interview protocol refinement framework was created, for such a protocol it is important that it promotes the conversation but still answers the research question (Castillo-Montoya, 2016). An interview Protocol Matrix was created containing the minimal requirements setup by Castillo-Montoya (2016), this was extended to also cover the major categories from the literature research. This interview protocol got peer reviewed and improved based on feedback provided. The final interview protocol is added to appendix 6.2.

3.3. Data analysis

With the data gathered it might be possible to enrich the list of C&SF's generated from the literature study with new not yet covered C&SF's or find relations between them. There is a requirement to go back and forth between earlier data and re-analysis and refine them on newly collected data. Saunders et al. (2016) advice to use the Thematic Analysis procedure for that. The recorded interviews will be transcribed keeping special attention in the tone and or emotions being expressed by the people interviewed since these could easily be missed (Saunders et al., 2016). Specialized software "F4analyse" will be used for transcribing and coding of all the interviews.

Because this is purely a deductive approach Saunders et al. (2016), mentions that it is possible to use 'a priori' codes extracted from the literature study, with the side note that this may lead towards having to revise the prior code list if it is inadequate. For this research the six categories will be used as 'a priori' codes, in which they also serve as a kind of a theme function. During the process of coding these will be enriched, subdivided, removed or merged. This process of coding will allow the linking of parts of data that refer to the same category or theme, which will allow the comparison of the results and improve the analysis. After every interview, the transcription will be done, and the data will be coded accordingly. During the analysis coding of themes and relationships between them will be looked for and refined when needed. Once all the interviews are coded an overall analysis will be done on all findings to better understand themes and relationships across all results (Saunders et al., 2016).

3.4. Reflection w.r.t. validity, reliability, and ethical aspects

This research will have its limitations with regards to its external validity and reliability:

Construct validity
There will be multiple sources of evidence, like multiple interviews per company, which will improve the overall validity of evidence provided. While collecting all the data a chain of evidence will be generated, which is also relevant for the data collection it selves (Yin, 2017).

Internal validity
This is an exploratory research and because of that internal validity is not applicable (Yin, 2017).

External validity
The amount of "scientific" top performing companies being interviewed for this research is relativity small (three companies), this might have impact on the external validity. The C&SF's from this research can still be considered as valuable for the external validity, because the information from these three companies enriches the big set of C&SF's based on the scientific research, which might allow the generalization of this for similar "scientific" top performing companies or companies who are facing the same C&SF's in the real world.

Reliability
To ensure reliability on what "data science" is everyone will be asked to describe that, this will prevent socially desirable answers, next to that only subject matter experts will be interviewed. It is important to do multiple interviews per company this will ensure the trustworthiness and reliability. To further ensure the consistency and replication of this research an audit trail will be kept while collecting data, which will enhance the reliability of this research. Via the use of the Interview protocol refinement method improvements to for the reliability and quality of data extracted from the interviews can be ensured (Castillo-Montoya, 2016).

Ethical aspects
People will be interviewed for this research ,which means a written and or verbal acceptance for participating is required. The interviewees need to be aware that this interview is fully voluntary and can stop the interview without any reason. They will be asked if recording these interviews is accepted and that answer will also be in the recordings. The full interview protocol will be explained before the actual start of the interview, which includes a full explanation about what the interview is about, for whom it is meant for, and that the real intent is to publish this thesis, all of this will also be in the recording.

4. Findings

The three companies that were selected needed to work on a day-to-day basis DSP's (minimum ten projects per year) and work according to a project methodology (scrum / waterfall / etc.).
For each company, an executive manager, business translator / project manager, and data scientist were interviewed. These three companies will be described in more detail below with some examples of projects they do and their approach to DSP's.
The information gained from the interviews has been used to enrich these case descriptions.

4.1. Case description of company "A"

Introduction company "A"

Company "A" is a big Dutch Business to Business (B2B) multinational company active in 100+ countries with more than 23,000 employees and makes components for products almost everybody uses / consumes in their daily life.

Table 2 | Company "A" short DSP summary

Size of the Company	Active in what type of industry('s)	Years active in Data Science Projects	Amount of unique roles involved in DSP's.
23,000+ employees	Health and materials	20	7

Strategy on DSP's

The company has in its corporate strategy "digital" as an enabler, which has led the company into a digital transformation that supports their three value drivers: customer intimacy, operation efficiency and new business models. The digital strategy requires the company to accelerate and scale data analytics and automation or Artificial Intelligence (AI).

Organizational setup to support DSP's

This company is a decentralized organization, however to better serve their internal customer/business they have setup a temporary Data Analytics Centre of Excellence (CoE) team to lead and orchestrate the company's endeavours in becoming an insight-driven organisation. This team should cease to exist once successful and when the internal customer/business are able to perform it on their own. Most of the data scientists, which work for this company, are based within the Research and Development (R&D) department, which makes them available for the whole company.
Organizing data science and its related projects was a challenge at first as company "A" said: *"we didn't know how to organize or orchestrate and that's why it was defined that we needed to set up a data centre of excellence to really pull this together to actually start delivering. And there it was also decided that it's actually a temporary team because basically our goal is to put this back into the organization so that everybody should be able to do this.".*

Mindset on DSP's

Company "A" says, "I think it is changing tremendously. I think now the business sees data as an enabler. Data science projects is something for which data science skills are needed". With that, they are solving nonlinear complex processes and are developing either machine learning scripts or other AI technologies to solve complex problems. It can be seen as an overarching component that includes all sorts of AI, machine learning, statistics, and some sort of advanced analytics, which goes from descriptive to predictive to prescriptive.

Maturity on DSP's

In terms of the maturity level of doing DSP's, the interviewed members of the company consider the company as experienced; however in terms of discovering the benefit of machine learning, they state that their experience is less than three years.

DSP's example

A DSP that they are currently working on is about a fermentation factory, which produces a product. During the production cycle there are different steps. They observe that in four of the six fermenters unexplainable variances are occurring in their process and performance. On Monday "Fermenter A" can be the best performing fermenter, on Tuesday "Fermenter B" and on Thursday "Fermenter C". Meaning that there is a lot of variation in the output per fermenter, but also there is a lot of variation in the total output of the six fermenters. The factory is outfitted with sensors, which gather data every minute. A team of data engineers/ scientists and SME's are trying to find if they can predict (via the use of a Random forest data model) the output based on these sensors and features derived from these sensors. Using ML interpretability tools such as "SHapley Additive exPlanations" (SHAP), feature importance, the team was able to identify features that influence the variance in the process and adjust these features to improve the performance.

DSP's approach

The approach of Company "A" of doing DSP's is based on their own developed 3 steps approach; the "Diagnostic", the "Minimal Viable Product" (MVP) and the "Scaled-up solution". After every step, there is a validation to continue or to cancel the initiative in close collaboration between the business and the data analytics CoE team. In the diagnostic step, the first thing done is a value and feasibility assessment of the project. Then they perform a project intake, which delivers a value case one slider and start designing the advanced analytics solution, which takes into account the following: the data availability, the previously done assessment, proposed project approach, a paper mock up solution and the recommended project team setup. With all of that they execute the prototype of the solution. If everything is good, they go into an MVP phase where they deliver the value in iterations as fast as possible. Finally, they go to a Scaled-up solution and make sure the requesting business is trained on how to use it and the handover to the run and maintain organization for solution refinement.

4.2. Case description of company "B"

Introduction company "B"

Company "B" is a small Local Dutch company specialized in marketing and data science, which mostly works for big international B2B companies. There they do strategic consulting but also just general statistics.

Table 3 | Company "B" short DSP summary

Size of the Company	Active in what type of industry('s)	Years active in Data Science Projects	Amount of unique roles involved in DSP's.
13 employees	Consultancy working for chemical, logistics and insurance companies	10	7

Strategy on DSP's
Data science is one of their main pillars and it has a clear relationship with their strategy. That means that some problems (which for example they could not solve three years ago because of lack of knowledge or computing power) they can go back towards and pick those up and support their business and themselves with those tactics.

Organisational setup to support DSP's
The organization of company "B" is equipped for the DSP's they are currently running. In their hiring process they ensure that everyone is a data scientist at its core and they can also do additional role like the data engineering part or the model redeployment. This provides the company additional flexibility. In teams they can combine certain persons with each other to best fit the purpose. If they have a very complex integration data-wise, they put the data scientist on who has more data engineering experience. If they have a very complex model, they take the one who has a very mathematical background and so on.

Mindset on DSP's
For company "B" the mindset of data science is their core business. However, they notice that their clients mindsets are changing tremendously; they see data as an enabler. Company "B" consider DSP's as any business task whose goal can be defined in terms of data. Almost all the work they do is related to something in data science. But doing DSP's takes time and time is an issue, as Company "B" explains: *"That's usually big pitfall, resources is not issue, money also not, but doing the whole data science project in three months is a bit ambitious. There's not enough time made available. They should not treat it as a traditional project with the fixed pieces in time, because these projects are usually a loop. You model, you find out something is wrong with it, you go back to the data again, etc. it takes time. I think at least every company should double their project time. I am not saying project spend but project time.*

Maturity on DSP's
Data science is their core, to quote, "that's our core, a strength. We get this because we have expertise in data science and connecting a business problem to data science. So yeah, that is what makes our company Unique. We do this work already more than 10 years, and back then it was called mathematics. So, it is more of a perception on what you call data science, I think".

DSP's example
An example of a DSP that they have done, involved the improvement of operational efficiency of food grade containers. They saw that if you have wine that needs to be shipped and there is a wine season in Europe and there is a wine season in South America, that those are completely the opposite from each other. So there are companies with a container park on both sides of the world being idle half of the year. What their data science model found out was that you can actually continuously use the same containers while shipping halfway into the season to the other side of the world. Reducing the idle time of the food grade containers.

DSP's approach
The approach of Company "B" of doing DSP's is according to the Team Data Science Process (TDSP) of Microsoft. "Data science projects are an agile, iterative data science methodology to deliver predictive analytics solutions and intelligent applications efficiently. TDSP helps improve team collaboration and learning by suggesting how team roles work best together. TDSP includes best practices and structures from Microsoft and other industry leaders to help toward successful implementation of data science initiatives. The goal is to help companies fully realize the benefits of their analytics program." All the steps are described in appendix 6.3.

4.3. Case description of company "C"

Introduction company "C"

Company "C" is a big US multinational company active in 150+ countries and manufactures healthcare products and related services.

Table 4 | Company "C" short DSP summary

Size of the Company	Active in what type of industry('s)	Years active in Data Science Projects	Amount of unique roles involved in DSP's.
90,000+ employees	Health	10	9

Strategy on DSP's

Information has become part of the company's global strategy as it needs to support on a global scale overall big strategy topics like "being efficient as an organization", "reducing overall costs" and "improving operational excellence". They are building a core "foundation of data", which should allow the company to link all the data assets they have and gain insights from that in a democratized manner. This needs to be supported by an Analytics Academy to improve the knowledge of the company on how to benefit from data science. The company has setup a specialized department which is taking care of all that and has the responsibility to break the paradox of being a technology-driven company and becoming an information-driven company.

Organisational setup to support DSP's

This company is a decentralized organization, which has multiple teams of data scientists spread across its Business Groups (BG) or as they say, *"So if I look at data science and its controlling, that's now scattered all over the place. So, I would say our company does not have a centre of excellence, but we have pockets of excellence.".* As previously mentioned, they have setup a specialized department called "Central Information Management Services" with its main purpose as coaching, driving, and making sure data science is successful.

Mindset on DSP's

According to them, data science is more classified as an advanced analytics area, more on the AI Machine learning production kind of space. The mindset as described by company "C":
"I think the business still need to get better understanding of it. So, one of the initiatives that we do on a global level is building an Analytics Academy to actually start training businesspeople. But especially also senior executive people in the organization to increase what we call the data literacy of the organization."

Maturity on DSP's

Concerning their maturity, they consider themselves as "still learning" as an organisation to understand how to best execute DSP's. As they say, "*a data science project is not like a software project where you can finish it with the timeline and say yes, we can do it and things like that. It is a very iterative process. Sometimes you might get the results, sometimes you might not get the results. Things like that".*

DSP's example

As an example of a DSP that, they are doing right now involves inventory, which is a huge problem for them. At any given moment in time, there is about 3,5 billion dollars worth of inventory and at the same time they have almost billions of dollars in back orders. So, on one end they have products gathering dust while on the other hand they have accounts that cannot be satisfied. With the use of

data science, they are investigating how they can reduce those backorders and reduce their inventory as well.

DSP approach

The approach of Company "C" of doing DSP's is to start with identifying, which use cases benefit the company the most and can they be mapped towards what they call "buckets" of opportunities like for example "operational excellence". Once the use cases are mapped, workshops will be organized with the stakeholder groups, in which these use cases can be prioritized. As soon as that is done, they will start the selected use case as small as possible and go for an MVP, and if that fits the expectations they continue. Next, they build a "quick and dirty" first iteration model to see if they can already prove that this use case has real value for the business. Once the business is convinced, the next step is to optimize the model in an iterative approach until they can automate the model in such way that there is no more manual intervention needed. In that final step they will fully operationalize the data science solution.

Table 5 | Case studies similarities and differences

Similarities and differences	**Company "A"**	**Company "B"**	**Company "C"**
Size of the Company	23,000 employees	13 employees	90,000+ employees
Active in what type of industry('s)	Health and materials	Consultancy working for chemical, logistics and insurance companies	Health
Years active in Data Science Projects	20	10	10
Amount of unique roles involved in DSP's.	7	7	9
Strategy on DSP's	The company has in their corporate strategy "digital" as an enabler, in which they identified the need for DSP's.	Data science is one of their main strategy pillars.	The company wants to break the paradox of being a technology-driven company and becoming an information-driven company.
Organisational setup to support DSP's	The company is a decentralized organisation, but has a specialized department to make DSP's successful.	Their organization is fully focused and equipped for the DSP's they are currently running.	The company is a decentralized organisation, but has a specialized department to make DSP's successful.
Mindset on DSP's	They are solving nonlinear complex processes and are developing either machine learning scripts or other AI technologies to solve complex problems.	They consider DSP's as any business task whose goal can be defined in terms of data.	According to them data science is more classified as an advanced analytics area.
Maturity on DSP's	Consider themselves as experienced, however in terms of discovering the benefit of machine learning they state that their experience is less than three years.	Consider themselves well-experienced and see DSP as their core strength.	Consider themselves as "still learning".
Data Science project approach	They have developed their own 3-step approach: the "Diagnostic", the "Minimal Viable Product" (MVP) and the "Scaled-up solution" for doing DSP's.	Fully structured according to the TDSP approach of Microsoft.	A BG driven approach in which DSP's are done in small steps with loops back to the business to verify the expectations.

Use case assessment	Orchestrated and controlled by a central temporary "Data Analytics CoE" team to ensure success and facilitate required resources. Prioritization of projects is done by this team and business supports them.	Gets hired at the stage that this is almost determined or gets requested to assist / advice.	Facilitated and driven by a "Central Information Management Services" to help make DSP's a success. Prioritization of projects is done by the business, and the team supporting them.
Data science resources	Corporate pool of experts available for all BG's with some specialisms for certain types of DSP's.	Corporate pool of experts available for all. On a multi-skilled level to allow flexibility across DSP's.	Experts are located within the BG's and have their own maturity level and expertise related to the needs of those BG's.
Main focus of the specialized data science team.	Focus on guidance and orchestration of data science advanced analytics within the company.	N/A	Focus on facilitating the BG's data science teams / community to work in one way on one platform to allow reuse of models/data/etc.

4.4. Cross case findings

The findings are presented in the six categories in which the C&SF's for these companies are analysed.

Value

Data science projects (DSP) have a certain value for business, may it be in actual EBITDA profit or just operational efficiency improvements. There is a lot to be achieved by doing DSP's. It is essential to identify the value from the beginning, measure the value at each of the steps, and have the business owner always engaged and validate this. If you talk about value generation, it has to move the "needle" for someone and that can be a move for the complete organization or just a customer. At company "A" they say: *"I am still daily amazed how much value we can still unlock if we work in a different way and then using tools and techniques like data science."* However, they also mention*" The value for data science is basically the outcome and you will not get the outcome, unless you productize all your insights. Therefore, you productize your data science and embedding it into either dashboards, or an application or an RPA or whatever you know as a product.".*

Doing DSP's alone is not enough, its needs to be embedded in the organization. Change management is the key to success because the value that you will get from any data science model will need to be used by the end users. At company "B" they say *"A project doesn't end when a project ends. Especially for data science related topics, it is important to maintain it. I think a lot of times companies forget that and think, oh well, we now built a model. Then after three months will that same model still be correct? No, they need to keep checking it or they need to have some ways of validating it.".*

Table 6 | Category "Value" specific C&SF's

C&SF's	**Company "A"**	**Company "B"**	**Company "C"**
The importance of being able to correctly identify the Value of DSP's for the company.	Applicable	Applicable	Applicable
Having the ability to embed data science results into the company way of working.	Applicable	Applicable	Applicable

People

There are several roles involved in doing DSP's. The three most reoccurring mentioning of roles within these case studies are the data engineer, business translators and the data scientists. Not only it is important to hire employees with these skills but also continuously develop and maintain their analytical and technical skills. This can be achieved via different methods. One could be as easy as a *"brown paper bag sessions"* once a week where they share their knowledge and skills as done by company "B". Or the more extensive solutions of company "A" and "C" were they have or are setting up a sort of an Analytics Academy to educate not only the data science community but improve the knowledge of data analytics on all the levels within the whole company.

When setting up a DSP team, it is advised by all the companies that it should be a multidisciplinary team as company "A" describes it: *"the more diverse team it is, the better you can solve the problems, but you will have the challenge to manage the team in an efficient way.".* SME's are key in doing DSP's because they know how the business is operating and can create context between the data science part and the business part.

The importance of having knowledgeable people in the team as stated by company "A": *"it's easier to build a data science capability within the SME that you have in your organization than to hire a data scientist and make them learn the SME expertise of your business. Today in the market there are lots of platforms that have already plug and play machine learning algorithms that you can run. The understanding of the algorithm is not anymore very important. What is important is what data you use as input and how to read the output from the inside, that's where the SME comes in".*

Company "C" adds: *"With setting up data science kind of projects or organizations its natural that you go for the technology kind of people at first. I think that what you need to do first is actually to get businesspeople to understand the technology. Also remember a data scientist does not have to be in your office. We are also talking to each other at this point it does not really matter whether we are. 20 kilometres away from each other or 2000 kilometres away from each other".*

Table 7 | Category "People" specific C&SF's

C&SF's	**Company "A"**	**Company "B"**	**Company "C"**
Having the correct roles like Business translators, Data Engineers, and data scientists within the company.	Applicable	Applicable (they combine multiple roles in one)	Applicable
Bringing and keeping the employees of your company on a high enough level to understand and support DSP's.	Applicable (via an "Data Analytics Academy")	Applicable (via company organized "brown paper" sessions)	Applicable (via an "Analytics Academy")
Making sure that you have multidisciplinary diverse teams doing DSP's.	Applicable	Applicable	Applicable
Importance of having people with actual business knowledge (SME's) vs data science knowledge in DSP's.	Applicable	Applicable	Applicable

Technology

From the technology category there were not a lot of C&SF's. The companies focused on having the right tools to do data analytics. They also recommend that if you want to be doing DSP's in a reusable way you have to approach it from a central technology platform. As company "C" says: *"I mentioned before when we talk about technology is etc. Make sure that whatever you going to do is going to be a building block for anybody out there."* Technology is considered as a real enabler which can speed up DSP's but it mostly depends on who uses the technology instead of the technology it self.

Table 8 | Category "Technology" specific C&SF's

C&SF's	Company "A"	Company "B"	Company "C"
Making sure to have a central technology platform for DSP's.	Applicable	Applicable	Applicable

Data

Data governance is important in all of the case studies in ensuring that there is focus on data quality. According to Company "C": *"So data governance is a key topic, including data quality. Recently we also introduced a data catalogue kind of function to actually start documenting and creating dictionaries around the data that we managed centrally. From a control perspective, but also from an end user perspective so that we can then train end-users to actually start using what we call a curated dataset. So that we have centralized views of data out there where we want to push out these datasets or actually bring users to these datasets to reuse the similar dataset instead of creating multiple solutions for every individual user.".* According to company "A" Data governance is difficult to apply: *"It depends at which stage of maturity you are within the organization. But regardless, I think every company will struggle and suffer from what I call a data governance syndrome when it comes to advance analytics."*

Table 9 | Category "Data" specific C&SF's

C&SF's	Company "A"	Company "B"	Company "C"
Importance of Data Governance	Applicable	Applicable	Applicable

Process

A big difference found while comparing the case studies is their approach on the "use case assessment". At Company "A" they have a specialized team which challenges the business on the use cases and their potential value and if they will accept that use case as a DSP. Company "B" normally skips this part because they are a company which can get hired by clients like "A" and "C". They join when there is a need defined and assist those companies from there onwards and look backwards to get the business understanding of why that project was chosen. Company "C" uses their specialized team to facilitate and let the business decides if a use case becomes a DSP via their "Use case mapping and prioritizing".

It is noticeable that at all the three companies do a lot of projects in an agile scrum methodology kind of manner, even though they all mention that they don't want to be to strict on any methodology. Company "B" has a strong preference to use *"The Team Data Science Process (TDSP),* which they consider their structured framework for doing data science projects.

To make sure that all the output and insights created by these data science projects are not lost they use several solutions. At company "A" they use a model library where they can check the health of the models created and make sure that they keep running as they should be. They have setup a team within the IT department to also ensure a similar part for the data engineering work. Both companies' "A" and "B" are using solutions like GitHub or Bitbucket which allow for version control of code and models which was created for both the data engineering and data science work.

Company "C" actually had invested in new technologies, and now have what they call a data science environment where this can be done on a production system.".

Table 10 | Category "Process" specific C&SF's

C&SF's	Company "A"	Company "B"	Company "C"
Doing DSP's in an agile scrum methodology kind of manner	Applicable (but not enforced)	Applicable	Applicable (but not enforced)
Importance of the reusability of use cases and or insights created by DSP's	Applicable	Applicable	Applicable

Organization

The most mentioned C&SF's were related to a lack of organizational acceptance / adoption of data science on all levels within the company. This mostly had to do with the fact that data science is not always correctly understood within the company and when they do understand, the willingness to accept the results became the next challenge. As company "A" nicely says *"being right versus being perceived to be right. I can pinpoint that there is two million lying on the floor, but you have to tell the story in such a way that they are actually willing to bend over and pick up the money. So, perseverance is needed because, you're a helping the company to work in a different way."*.

What they found is that engaging business or even have them join the DSP's or parts of those processes helps in acceptance. During the interviews it became clear there is a balance in knowledge to be achieved by both parties within the company. To quote company "B" on this importance "*I think its really important to engage business into the end product so that they know what's being built for them and they know what is inside the data model, and they also have the feeling that they are engaged with the whole process, so they have a sort of ownership with that as well.*" Company "C" added "*collaboration between business and IT is very crucial. We see that this team is successful because the IT Subject Matter Expert (SME), is really close to a business SME and they really make this work. They have daily meetings with each other in which they really discuss the outcomes.*" Full involvement can be difficult for business because often these projects are put on top of their daily operations and so create resource issues. Getting senior management support on this means that they really want to have continued improvements to generate value. That will reduce the business pressure on time and money. Just doing the DSP's is not enough, the output should be used as previously mentioned only then you can be successful with data science or as company "A" says: "*If you consider success being that you as a company profit, you need to walk the last mile, meaning you need to implement the findings out of the data science exercise.*".
The following table shows a summary of the most important C&SF's and if they are applicable for all cases.

Table 11 | Category "Organization" specific C&SF's

C&SF's	Company "A"	Company "B"	Company "C"
The Importance of having the Business (including SME's) understand what data science entails.	Applicable	Applicable	Applicable
Importance of getting Senior management support on DSP's.	Applicable	Applicable (they even find it even more important to have middle management support)	Applicable
Engaging business (SME's) or even have them integrated into the DSP's.	Applicable	Applicable	Applicable
Having the correct people and approach on doing DSP's.	Applicable	Applicable	Applicable
Having organizational acceptance / adoption of DSP's.	Applicable	Applicable	Applicable

Relation across the C&SF's

During the analysis of these results it became clear that there was a relationship with a reoccurring element that had a dependency influence or accelerator capability towards other C&SF's. This was having a SME join and actively participate in a DSP. It has positive effects (accelerator, dependency) on the success of that DSP. In table "12" you can see what the effect was of SME active participations in DSP for the case studies including the why and how that effect works for the C&SF's identified as important in this research.

Table 12 | Positive effects of having an SME in a DSP for 6 relational C&SF's

No.	C&SF's	The Effect of an SME joining and actively participate in a DSP (the WHY)	Type of effect (the HOW)
1	Ability to correctly identify the Value of DSP's for the company.	The SME knows where the improvements can best help the company and help correctly identify the Value of a DSP.	***Accelerator***, the SME can help with identifying the value of a DSP with his level of expertise from within the level on which the DSP is going to be initiated.
2	Ensuring acceptance / adoption of DSP's.	The SME enforces the acceptance /adoption because it will improve his own work.	***Dependency***, without SME involvement the results of an DSP will not easily get accepted / adopted.
3	Ability to embed data science results into the company way of working.	The SME is helped a lot with the outcomes of a DSP and can use that to improve his own work.	***Dependency***, without SME involvement the adjustments needed to make the results work within the company will not be implemented.
4	Business understanding what data science entails.	By involving the SME in the DSP, he/she learns what data science entails from first-hand.	***Accelerator,*** the SME can share his knowledge and show the impact of data science directly from his/her business perspective.
5	Improvement in data quality and availability within the company.	The SME is the main consumer and or creator of the data, which is used in DSP's. While participating in such a DSP the pain of bad data quality or the lack of its availability is something they can help solve and then will also understand the importance of having that corrected for the future.	***Accelerator,*** the SME can easily pinpoint what needs to be adjusted to improve the data quality and availability because he/she knows how this works from their business perspective.
6	Bringing and keeping the employees of your company on a high enough level to understand and support DSP's.	The SME's will learn from participating in a DSP's and will share that knowledge through the rest of their team. This will cause a knowledge demand that forces the organization to supply (As was seen in the results that companies "A" and "C" are setting up Analytical Academies. Some are even building in data science capabilities in their SME instead of hiring people).	***Accelerator,*** The SME are needed for the acceptance and implementation of the results of a DSP. So, knowledge transfer of what a DSP would have delivered is needed. Taking them along in the process accelerates the understanding and knowledge.

The SME has a leading role but is not the only thing that ensures the success for DSP's. In Figure 2 there is a relation diagram of all the 6 SME related C&SF's (ovals) as mentioned in table 12 (the numbers correspond with that table). These C&SF's have a relationship with not only the SME, but also with other mitigations done by the case studies to improve the success of DSP's as were mentioned in the previous subchapters. The bigger the line between C&SF's and the other elements the bigger the influence.

Figure 2 | The relations between the C&SF's (ovals) (see table 12) and the "solutions" applied by the case studies.

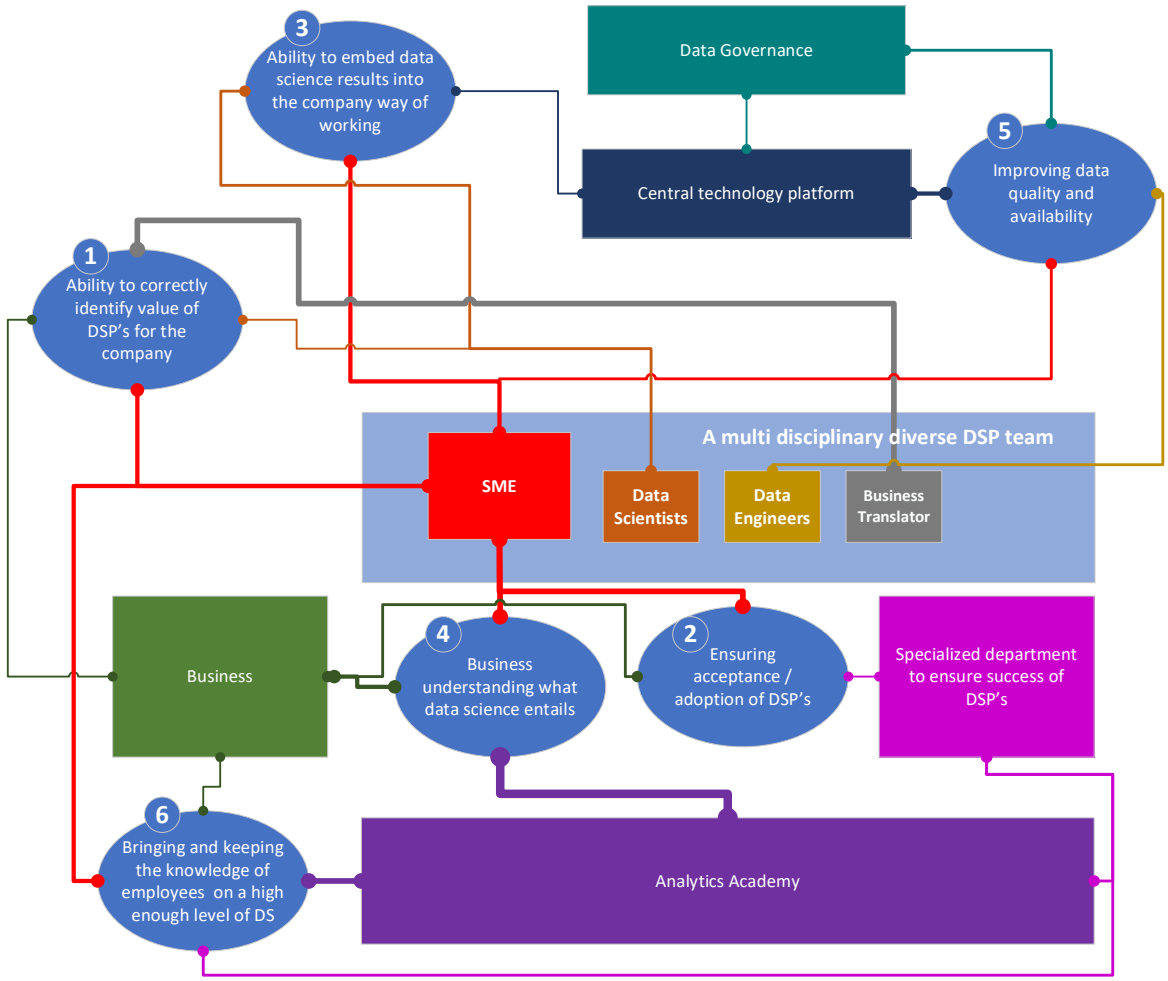


Table 13 | Supporting table to figure 2

No.	C&SF's	"solutions" applied by the case studies
1	Ability to correctly identify the Value of DSP's for the company.	SME is needed together with Business translator and Business to correctly identify value of DSP's.
2	Ensuring acceptance / adoption of DSP's.	SME's together with Business and Specialized department are assisting in the overall acceptance/adoption of DSP's in the company.
3	Ability to embed data science results into the company way of working.	SME's and Data scientists are able to get the results with the help of a central technology platform embedded in the company way of working.
4	Business understanding what data science entails.	SME's and Business can increased their understanding via the Analytics Academy, and the SME also via its active participation in the multi-disciplinary diverse DSP team.
5	Improvement in data quality and availability within the company.	The SME's are at the data source of what is needed for DSP's and with that can influence its data quality and availability, via the help of the Data engineers and Data governance they are able to improve this and have it available on a central technology platform.
6	Bringing and keeping the employees of your company on a high enough level to understand and support DSP's.	SME's gain knowledge during their active participation in DSP's which will contribute to this C&SF next to that there is the knowledge that can be gained from the Analytics Academy and the expert knowledge and guidance provided by the specialized department, which all contribute to the knowledge of all the employees including business.

5. Discussion, conclusion, and recommendations

5.1. Discussion - reflection

Key research finding

What can be concluded from this research is that there were some significant differences between the case studies but also similarities in solving their C&SF's. However, it all led to the most important C&SF's, which is the importance of having a SME actively participate in the DSP's. A lot of the results from chapter 4 are linked to SME participation in DSP's, which causes a chain reaction on the C&SF's of DSP's. The SME's have a kind of accelerator role in DSP's. The active participation will show case the actual value of DSP's because there is acceptance and a clear need from the business to have these DSP's done, mainly because the SME can pinpoint where exactly the issue is. There is also a dependency on the SME because he/she is heavily involved making sure that the solution will be embedded in the organization because this has a clear benefit for the SME's own work and/or achievements. The importance of data quality is no longer a discussion because the SME understands the importance and has the power to make adjustments within that. Lastly, but surely, it will increase the knowledge of the organization in regards to data science since the SME's are fully involved.

In the cross case findings it became visible that there are other components which also contributed to the success of DSP's next to the SME's active participation. Ensuring that the company has proper roles defined within a multi-disciplinary diverse DSP team helps in the ability to correctly identify the value of the DSP's for the company; it assists in the ability to embed data science results into the company's way of working and the overall speed and the success of DSP's. Ensuring that the knowledge within the company on Data Science stays on a high enough level and business understands about what data science improves are also being addressed via Analytical Academies and the specialized departments. These specialized departments also help with ensuring the acceptance and adoption of DSP's and guide the Analytical Academies. Having a central technology platform can assist with improving the data quality and availability and will improve the embedding of the data science results into the company's way of working and speeds up DSP's in general. All of these relational links are visualized in figure 2 in chapter 4.

Key research finding versus Literature

When comparing the "key research finding" versus the important categories from the literature it is noticeable that there is something missing. According to the literature it is important to have clear expectations of how to setup a DSP, discover the value, the business case and how to do the project from a process perspective (Dutta & Bose, 2015; Koronios, 2015; LaValle et al., 2011; Schüritz et al., 2017; Schwarz et al., 2014; Viaene, 2011; Vidgen et al., 2017). The literature doesn't provide an overview of relations between C&SF's. It also has not identified the relationship of these C&SF's with the active participation of SME in DSP's. This will greatly reduce the difficulties of having the DSP's results like the data model and/or insights reincorporated back into the organization, not only from a technical point of view, but more importantly is the acceptance/adoption of the internal organization on these DSP results.

The results versus Literature

What was remarkable is that of the six major categories "technology" and "data" are not the key drivers according to the case studies. In short, there is a need have these things arranged, but having people able to use and interpret them is far more important than having the top of the line technologies or all data you can get your hands on. With the correct "knowledgeable" people and an organization that understands how data science works a company can do great things with data science to improve it self. The literature recommends that companies, which are planning to do

DSP's should be setup from an independent , centralized analytic unit (Koronios, 2015; LaValle et al., 2011; Schüritz et al., 2017). These cases studies acknowledged that they had a lack in this data science knowledge so they setup specialized teams exactly as the literature recommended, in order to show case the value and importance of data science for the company and help the rest of the company to be successful with that as well. To ensure that this knowledge can spread more widely, these teams help their organizations via setting up so called "academies" within the companies in which its employees can be trained and tracked on their knowledge of data science and analytics. The literature also advises that these teams should focus on minimizing the knowledge gap but has no hard recommendations how to do so (LaValle et al., 2011; Obasa et al., 2017; Schwarz et al., 2014; Viaene, 2011; Vidgen et al., 2017).
From the literature there were several recommendations from a technology point of view on what kind of tools are best to use etc. However, from the interviews there was only one real C&SF linked to technology, and that was to have a central platform available and what ever got created on the platform is shareable and reusable.

Data was one of the most mentioned categories from within the literature, however from this research the amount of C&SF's on that topic were not leading. From the data category, the biggest concern was data quality and next to that is data access, which is in-line with the expectations (Koronios, 2015; LaValle et al., 2011; Philip Chen & Zhang, 2014; Schüritz et al., 2017; Viaene, 2011; Vidgen et al., 2017). However the technical requirements like migration, integration, security or even privacy were not key for these companies, which was expected to be important (Koronios, 2015; Schüritz et al., 2017; Schwarz et al., 2014; Viaene, 2011; Vidgen et al., 2017).

From a process point of view there is a deviation according to the expectations from the literature. According to the literature, it is advised to have a well-defined project management setup following a clear methodology (Koronios, 2015; Saltz et al., 2017; Viaene, 2011; Vidgen et al., 2017). If we look at how the case studies are actually doing it, only one company holds true to a well-defined project management setup while the others consider a more flexible way of working.

Reflection
Now I do not know if these new insights or if these research findings state the obvious and this is just a logical sequence in the chain of doing DSP's. My literature research was very broad and touched a lot of elements related to data science, which caused the overall approach to be broad as well even though I tried to mitigate that with refining the interview protocol. Consequently, it lacks depth on certain topics, and I noticed that mostly in the answers from the interviewee's on certain topics. I could have asked for more details in order to get a deeper understanding which might have been more interesting for this research.
It might have been more valuable for this research if for company "B" a similar decentralized big multinational company like "A" and "C" would have been interviewed, because that would have given this research more focus on such companies.
Because of COVID 19 the interviews did not go as planned, it lacked the "physical connection" between interviewer and interviewee, which might have impacted the interviews. This is because via video chat or phone it is possible that you might miss certain emotions or body language.
The trustworthiness and replicability of the research should be there, all the companies have been interviewed via the triangulation of sources, and a chain of evidence was created while doing this research. The credibility of the findings in this research is also covered via the triangulation of the sources per company and the overlap between certain answers across companies also confirm this credibility. Via the chain of evidence and the use of the specialized transcription and analysis software, the confirmability is secured because this provides an audit trail which highlights every step of the data analysis.

If another researcher would do the same interviews with the same interview protocol and the same analytical coding at the same companies, he/she should come very close to the same results.
From a validity point of view I only had three “scientific” top performing companies being interviewed. This had an effect on the external validity because this research can now only be useful for similar kinds of “scientific” top performing companies; I expect these companies might face similar challenges and can use the results of this research to assess their own situation.
Improvement is needed because even though there were three “scientific” top performing companies, the sizing of the companies were too different to make this research applicable for all “scientific” top performing companies. From a construct validity this is still in-line with the expectations: multiple resources of evidence are used per use case and there is the chain of evidence created while doing the research.
To reflect if my research is only applicable for these kind of case organisations, I think not because even though these companies have been working for more than 10 years in data science, some of their challenges are the same as for companies who are just starting with data science. The learnings, which these three companies have had can be applied to any company starting or already doing data science. Every company has its own type of SME’s, which are faced with their own issues and challenges that data science could solve.

5.2. Conclusions

With this research I tried to get an answer for the main research question “*How can a company ensure the successful execution of a data science project (DSP)?*”. During the literature study it became clear that C&SF’s that are influencing the successfulness of DSP’s can be caused by numerous things. They all have their own level of influence on the success of a DSP. During the research, a lot of different relevant C&SF’s were found, which had a certain level of influence on the success of DSP’s for the case studies. Having a centralized analytical unit helping the company doing DSP’s is of great value and increases the success of DSP’s. Ensuring that the company is aware of what data science entails and how it can keep and or increase that knowledge is very important and a sort of company-wide training capability like an Analytics Academy helps in that matter. Having some sort of data governance and a central technology platform will improve the data quality and availability and allow for easier embedding of data science results into the company’s way of working. Making sure that the correct roles are defined within the company and have them all work in a multi-disciplinary diverse DSP team will increase its chance of success. The biggest finding is the relations of the SME’s active participation to DSP’s and the other C&SF’s including the contributing “solutions” which these case studies have applied. These relational findings across the six major categories provide new insights not yet described by any literature, which can be used by other companies to validate their way of doing DSP’s.

5.3. Recommendation for practice

The research provides practical advice other companies can apply when preparing their organisation for DSP. They can verify if they can relate to certain aspects of the case studies in this research and see if they can apply similar approaches.

Companies already involved in doing DSP can have a look at this research and see if they can take some of these learnings and perhaps make needed adjustments to improve the success of their DSP's.

For either scenario companies can investigate for themselves if the role of their SME's in their company is as profound as it was for these case studies and see what they can do with that information to improve the success of their DSP's.

5.4. Recommendation for further research

This research is by definition limited by scope and time of what could be accomplished. An important next step of such a research is the further research that can start based upon these findings.

Avoiding the possibility to take this importance for granted, similar research can be conducted in organizations that don't involve SME's in their DSP versus those who do, keeping in check that these organizations are based on equal level of data science maturity and size and type of companies.

What could have played a role in this research is the data science maturity level of the cases studies chosen. There was no official maturity level assessment done, and the rating of maturity in this research was based on the information provided by the interviewees. Having a data science maturity level measurement done with the case studies and have that compared to their successfulness of their DSP's could provide interesting results, which can be used to investigate if there are certain relationships to the findings of this research.

The scope of this research was only 3 case studies. To get a better understanding a broader research can be done, the scope should not only focus on the amount of DSP's a company had done but also the size of the company and whether the organization structure is centralized or decentralized as these factors can also impact their approach in doing DSP.

6. References

Alshenqeeti, H. (2014). Interviewing as a Data Collection Method: A Critical Review. *English Linguistics Research, 3*(1). doi:10.5430/elr.v3n1p39

Barratt, M., Choi, T. Y., & Li, M. (2011). Qualitative case studies in operations management: Trends, research outcomes, and future research implications. *Journal of Operations Management, 29*(4), 329-342. doi:10.1016/j.jom.2010.06.002

Castillo-Montoya, M. (2016). Preparing for interview research: the interview protocol refinement framework. *The Qualitative Report, 21*(5), 811.

The Data Science Process (TDSP). Retrieved from https://docs.microsoft.com/nl-nl/azure/machine-learning/team-data-science-process/overview#standardized-project-structure

Dutta, D., & Bose, I. (2015). Managing a Big Data project: The case of Ramco Cements Limited. *International Journal of Production Economics, 165*, 293-306. doi:10.1016/j.ijpe.2014.12.032

Knopf, J. W. (2006). Knopf Doing a Literature Review. *PS: Political Science and Politics, Vol. 39*(No. 1), pp. 127-132.

Koronios, A. (2015). Towards A Process View on Critical Success Factors in Big Data Analytics Projects In Proceedings of the Twenty-First Americas Conference on Information Systems.

LaValle, S., Lesser, E., Shockley, R., Hopkins, M. S., & Kruschwitz, N. (2011). Big Data, Analytics and the Path From Insights to Value. *MIT Sloan Management Review*.

Matsudaira, K. (2015). The science of managing data science. *Commun. ACM, 58*(6), 44-47. doi:10.1145/2745390

MGI, M. G. I. (2016). The Age of Analytics: Competing in a Data-Drive World. *McKinsey Global Institute (MGI)*.

Moya-Gómez, B., Salas-Olmedo, M. H., García-Palomares, J. C., & Gutiérrez, J. (2018). Dynamic Accessibility using Big Data: The Role of the Changing Conditions of Network Congestion and Destination Attractiveness. *Networks and Spatial Economics, 18*(2), 273-290. doi:10.1007/s11067-017-9348-z

Obasa, A. I., Salim, N., & Al-Taie, M. Z. (2017). Successful Data Science Projects: Lessons Learned from Kaggle Competition. *Kurdistan Journal of Applied Research, 2*(3), 40-49. doi:10.24017/science.2017.3.18

Okoli, C., & Schabram, K. (2010). A Guide to Conducting a Systematic Literature Review of Information Systems Research. *Sprouts: Working Papers on Information Systems, 10*(26).

Philip Chen, C. L., & Zhang, C.-Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences, 275*, 314-347. doi:10.1016/j.ins.2014.01.015

Provost, F. (2013). Data Science for Business - What You Need to Know About Data Mining and Data-Analytic Thinking. *Book*.

Saltz, J. S., Shamshurin, I., & Crowston, K. (2017). Comparing Data Science Project Management Methodologies via a Controlled Experiment Proceedings of the 50th Hawaii International Conference on System Sciences.

Saunders, M., Lewis, P., & Thornhill, A. (2016). *Research methods for business students*. Harlow; Munich: Pearson.

Schüritz, R., Brand, E., Satzger, G., & Bischhoffshausen, J. (2017). HOW TO CULTIVATE ANALYTICS CAPABILITIES WITHIN AN ORGANIZATION? – DESIGN AND TYPES OF ANALYTICS COMPETENCY CENTERS. *In Proceedings of the 25th European Conference on Information Systems (ECIS)*, pp. 389-404.

Schwarz, C., Schwarz, A., & Black, W. C. (2014). Examining the Impact of Multicollinearity in Discovering Higher-Order Factor Models. *Communications of the Association for Information Systems, 34*. doi:10.17705/1cais.03462

Stimmel, C. L. (2015). Building the Foundation for Data Analytics. *EDPACS, 52*(1), 1-13. doi:10.1080/07366981.2015.1059117

Viaene, S. (2011). The Secrets to Managing Business Analytics Project. *Article in MIT Sloan Management Review*.

Vidgen, R., Shaw, S., & Grant, D. B. (2017). Management challenges in creating value from business analytics. *European Journal of Operational Research, 261*(2), 626-639. doi:10.1016/j.ejor.2017.02.023

Walker, J. (2017). http://www.digitaljournal.com/tech-and-science/technology/big-data-strategies-disappoint-with-85-percent-failure-rate/article/508325. *Digital Journal*.

Yilmaz, K. (2013). Comparison of Quantitative and Qualitative Research Traditions: epistemological, theoretical, and methodological differences. *European Journal of Education, 48*(2), 311-325. doi:10.1111/ejed.12014

Yin, R. K. (2011). *Qualitative Research from Start to Finish*. New York: The Guilford Press.

Yin, R. K. (2017). *Case Study Research and Applications: Design and Methods*: SAGE Publications.

1. Appendix

1.1 42 Factors and 154 sub Factors per category with literature references

Category	No. Literature references	total sub Factors	Main Factors	Literature References
Value	7	11	Using analytics for improved decision making	Vidgen et all (2017), Steve Lavalle (2011), Schwarz (2014), Schuritz (2017), Chen (2014), Koronios (2015), Dutta (2015)
	2	3	Measuring customer value impact	Vidgen et all (2017), Koronios (2015)
	7	9	Establishing a business case	Vidgen et all (2017), Viaene (2011), Steve Lavalle (2011), Schwarz (2014), Schuritz (2017), Koronios (2015), Dutta (2015)
People	7	7	Building data skills in the organization	Vidgen et all (2017), Steve Lavalle (2011), Viaene (2011), Schuritz (2017), Obasa (2017), Koronios (2015), Dutta (2015)
	5	5	Analytics skills shortage	Vidgen et all (2017), Steve Lavalle (2011), Viaene (2011), Schwarz (2014), Koronios (2015)
	4	4	Technical skills shortage	Vidgen et all (2017), Viaene (2011), Obasa (2017), Koronios (2015)
	1	1	Acquisition and retention	Viaene (2011)
	3	6	Unit which does the analytics	Steve Lavalle (2011), Schuritz (2017), Koronios (2015)
Technology	6	6	Restrictions of existing IT platforms	Vidgen et all (2017), Schwarz (2014), Viaene (2011), Dutta (2015), Steve Lavalle (2011), Stimmel (2015)
	3	3	Managing data volumes	Vidgen et all (2017), Chen (2014), Koronios (2015)
	1	1	Visualization as storytelling	Viaene (2011)
	6	6	Managing a big data / analytics platform	Vidgen et all (2017), Stimmel (2015), Schwarz (2014), Schuritz (2017), Chen (2014), Koronios (2015)
Data	5	5	Managing data quality	Vidgen et all (2017), Viaene (2011), Steve Lavalle (2011) Chen (2014), Koronios (2015)
	4	5	Availability of data	Vidgen et all (2017), Steve Lavalle (2011), Chen (2014), Koronios (2015)
	5	5	Getting access to data sources	Vidgen et all (2017), Steve Lavalle (2011), Stimmel (2015), Schuritz (2017), Koronios (2015)
	3	3	Managing and integrating data structures	Vidgen et all (2017), Schuritz (2017), Koronios (2015)
	4	6	Managing data security and privacy	Vidgen et all (2017), Viaene (2011), Schwarz (2014), Koronios (2015)
	4	4	Data visualization	Vidgen et all (2017), Chen (2014), Koronios (2015), Dutta (2015)
	1	1	Defining what 'big' data is	Vidgen et all (2017)
	6	7	Managing data	Vidgen et all (2017), Stimel (2015), Viaene (2011), Schuritz (2017), Chen (2014), Obasa (2017)
Process	4	4	Producing credible analytics	Vidgen et all (2017), Steve Lavalle (2011), Chen (2014), Koronios (2015)
	1	1	Performance management	Vidgen et all (2017)
	4	5	Managing data processes	Vidgen et all (2017), Viaene (2011), Steve Lavalle (2011), Viaene (2011), Matsudaira (2015)

	2	2	Knowledge Management	Schuritz (2017), Koronios (2015)
	1	1	Manipulating data	Vidgen et all (2017)
	1	1	Ethics process	Viaene (2011)
	7	7	Project management	Vidgen et all (2017), Viaene (2011), Schuritz (2017), Saltz (2017), Matsudaira (2015), Koronios (2015), Dutta (2015)
	1	1	Broad adoption of analytics within organization	Schuritz (2017)
	1	1	IT governance for analytics	Schuritz (2017)
Organisation	3	3	Creating a big data and analytics strategy	Vidgen et all (2017), Viaene (2011), Koronios (2015)
	1	1	Defining the scope of analytics projects	Vidgen et all (2017)
	1	1	Legislative and regulatory compliance	Vidgen et all (2017)
	7	7	Building a corporate data culture	Vidgen et all (2017), Viaene (2011), Stimmel (2015), Schuritz (2017), Steve Lavalle (2011), Koronios (2015), Dutta (2015)
	6	6	Making time available for analytics	Vidgen et all (2017), Steve Lavalle (2011), Schwarz (2014), Schuritz (2017), Koronios (2015), Steve Lavalle (2011)
	1	1	Overcoming resistance to change	Vidgen et all (2017)
	2	2	Agreeing data ownership	Vidgen et all (2017), Steve Lavalle (2011)
	1	1	Managing costs of analytics	Vidgen et all (2017)
	1	1	Securing investment	Vidgen et all (2017)
	1	1	Using the data ethically	Vidgen et all (2017)
	1	1	Safeguarding reputation	Vidgen et all (2017)
	2	4	Working with academia	Vidgen et all (2017), Viaene (2011)
	4	4	Strong collaboration between IT and Business	Vidgen et all (2017), Steve Lavalle (2011), Schwarz (2014), Matsudaira (2015)

1.2 Literature research findings per reference:

Row Labels	Count of Literature
Vidgen et all 2017	35
Steve Lavalle, 2011	24
Viaene 2011	23
Koronios 2015	22
Schuritz, 2017	15
Chen 2014	8
Schwarz, 2014	8
Dutta 2015	7
Stimmel 2015	4
Matsudaira 2015	3
Obasa 2017	3
Stimel 2015	1
Saltz 2017	1
Grand Total	**154**

1.3 Interview Protocol

Question Number	Interview Questions	Research Question How can a company ensure the successful execution of a data science project?		Background information
		What are the minimal requirements for doing a DSP within a company (process, people (roles and skills), data, etc.)?	How can a company best set-up their organisation to do data science projects?	
	I would like to thank you once again for participating in my research by allowing this in the interview. As I have mentioned to you before, my study seeks to understand how a company can ensue the successful execution of a data science project. With the aim to create a basic guide which companies can use when they are doing or planning to do data science projects. **background info** Our interview today will last approximately one hour during which I will be asking you about your experiences, opinions and decisions made in regards to data science projects. For the record I would like to again ask again for consent and your permission to audio record our conversation. Are you still ok with me recording (or not) our conversation today? Thank you! Please let me know if at any point you want me to turn off the recorder or keep something you said off the record.			X
	I have setup the interview questions in such a way that I will cover several categories which are all related to data sciences these are: General questions, Value, Organisation, People, and Process related			X

	questions. In total there are 28 questions which are split over these categories which are between 3 and 8 questions.			
1	Could you please explain your role and its relation to data science?			X
2	What does your organization consider Data Science Projects?	X	X	X
3	How long is your company involved in data science projects?	X	X	X
4	How would you describe the experience your company has in doing data science projects?		X	X
5	Could you give an example of a data science project within your company?		X	X
6	What kind of challenges have you or are you experiencing during such projects?	X	X	
7	What are the crucial success factors for such projects?	X	X	
	Transition from general questions towards Value **Thank you for your responses so far I 's like to now ask you questions regarding the value data science projects have for your company.**			
8	How would you describe the value of data science projects within your company?		X	X
9	How does your company identify (Business cases) data science projects?	X		
10	if you could give advice for another organisation who is struggling in dealing with the value of data science projects, what would that be?		X	X

	Transition from Value towards Organization **Thank you for your responses so far. I would like to now ask you questions regarding organizing the ability to do data science projects.**			X
11	Can you describe if there is a relation between data science and your company's strategy?	X	X	X
12	How is your organization setup for data science project?	X		
13	How is data science perceived within the company?	X	X	
14	Could you describe the mindset of your organisation around how to deal with data?	X	X	X
15	Could you describe the top management thoughts about data science projects?	X	X	
16	How is "time" made available for doing data science projects?	X		
18	if you could give advice for another organisation who is struggling in dealing with the organizational setup and culture for managing data science projects, what would that be?		X	X
	Transition from Organisation towards People **Thank you for your responses so far. I would like to now ask you questions regarding People in data science projects**			X
17	What kind of roles do you have in your company supporting data science projects?	X		
19	How does your company ensure that the analytical and technical skills of its employees are of a high enough level?	X	X	
20	How are the units that are involved in data sciences projects setup, from a skill, domain, diversity level point of view?	X	X	
21	if you could give advice for another organisation who is struggling in dealing with the people side of things in relation to managing data science projects, what would that be?		X	X

	Transition from People towards Process **Thank you for your responses so far. I would like to now ask you questions regarding the "process" around data science projects**				X
22	Are there roles which are responsible for data and its quality, availability and accessibility? (no=25)				
23	Yes	Can you elaborate on this role and what are your experiences in regard to data and its quality, availability and accessibility	X	X	
24	No	How is your company dealing with the data and its quality, availability and accessibility?	X	X	X
25	How do you manage (project methodology) your data science projects?		X	X	X
26	How does your company make sure that what has been created and discovered is not lost?		X	X	X
27	if you could give advice for another organisation who is struggling in dealing with the process related elements of managing data science projects, what would that be?			X	X
28	Are there things which you consider extremely important for managing data science projects which are not yet discussed during this interview				X
	Thank interviewee and explain that you will keep them updated on the progress and that you will share with them the final thesis.				X

1.4 Team Data Science Process (TDSP)